COMPARATIVE EVALUATION OF ANNOVAR, OPENCRAVAT, AND NIRVANA WORKFLOWS FOR VARIANT ANNOTATION IN LARGE-SCALE WGS STUDIES

Asset Daniyarov[1,2], Rakhmetkazhi Bersimbaev[2], Ulykbek Kairov[1]

[1] Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, 010000, Kazakhstan
[2] Faculty of Natural Sciences, L.N. Gumilyev, Eurasian National University, Astana, 010008, Kazakhstan

**Corresponding author (s): asset.daniyarov@nu.edu.kz*

Background: Accurate variant annotation plays a crucial role in whole-genome sequencing (WGS) studies, enabling researchers to identify pathogenic variants, prioritize candidate genes, and better understand population-specific genomic diversity. Several widely used annotation tools, including ANNOVAR[1], OpenCRAVAT[2], and Nirvana[3] differ in supported databases, classification frameworks, and computational efficiency. However, systematic comparisons of these tools on large-scale population WGS data remain limited, particularly for underrepresented populations such as the Kazakh cohort.

Materials and methods: We compared the performance of ANNOVAR, OpenCRAVAT, and Nirvana using a high- coverage WGS dataset (~4.9M SNVs and InDels) from a Kazakh population sample. Each VCF file was processed independently with standardized database configurations relevant for clinical interpretation. All analyses were performed on an Illumina DRAGEN v4.3.13 server (Oracle Linux 8.9) equipped with 2× Intel Xeon Gold 6226R CPUs @ 2.90 GHz (64 threads), 512 GB RAM, and FPGA-based hardware acceleration. This unified computational environment ensured consistent performance measurements across tools. The comparison focused on the number of annotated variants, supported databases, detection of rare pathogenic variants, and computational runtime.

Results: Our analysis revealed notable differences between the tools. ANNOVAR annotated 4.93M variants in 2h 21m (4,391,599 SNPs, 539,696 INDELs) and reported 9 pathogenic and 4 likely pathogenic variants. OpenCRAVAT processed 4,913,713 variants in 7h 39m (4,376,625 SNPs, 537,088 INDELs), identifying 18 pathogenic and 13 likely pathogenic variants. Nirvana completed the analysis in 9m 21s, capturing 4,840,343 variants (3,907,526 SNPs, 932,817 INDELs) and reported substantially more clinically relevant findings, including 726 pathogenic and 277 likely pathogenic variants. Notably, ~22% of pathogenic variants were uniquely identified by a single tool, highlighting the complementarity of different annotation strategies (Figure 1).

Conclusion: No single annotation tool provides complete variant coverage. ANNOVAR offers speed and efficiency, OpenCRAVAT provides deeper predictive insights, and Nirvana enhances clinical interpretation, particularly for structural variants. Combining results from multiple workflows significantly improves annotation depth and clinical relevance, especially in large-scale population WGS studies.

Key words: whole-genome sequencing, variant annotation, ANNOVAR, OpenCRAVAT, Nirvana, population genomics

References:

1. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164-e164, (2010).

2. Pagel, K. A. et al. Integrated Informatics Analysis of Cancer-Related Variants. JCO Clin Cancer Inform 4, 310-317, (2020).

3. Stromberg, M. et al. in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics 596, (2017).

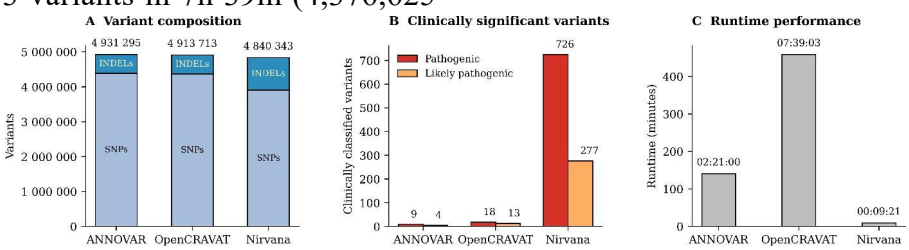

Figure 1. Comparison of variant annotation tools on WGS data